METHOD AND SYSTEM FOR PATH BUILDING IN A COMMUNICATIONS NETWORK

CROSS-REFERENCE TO RELATED APPLICATION(S)

[0001] This application claims the benefit of U.S. Provisional Application No. 60/287,069 entitled "METHOD FOR IMPLEMENTING A CLUSTER NETWORK FOR HIGH PERFORMANCE AND HIGH AVAILABILITY USING A FIBRE CHANNEL SWITCH FABRIC," filed April 27, 2001; U.S. Provisional Application No. 60/287,120 entitled "MULTI-PROTOCOL NETWORK FOR ENTERPRISE DATA CENTERS," filed April 27, 2001; U.S. Provisional Application No. 60/286,918 entitled "UNIFIED ENTERPRISE NETWORK SWITCH (UNEX) PRODUCT SPECIFICATION," filed April 27, 2001; U.S. Provisional Application No. 60/286,922 entitled "QUALITY OF SERVICE EXAMPLE," filed April 27, 2001; U.S. Provisional Application No. 60/287,081 entitled "COMMUNICATIONS MODEL," filed April 27, 2001; U.S. Provisional Application No. 60/287,075 entitled "UNIFORM ENTERPRISE NETWORK SYSTEM," filed April 27, 2001; U.S. Provisional Application No. 60/314,088 entitled "INTERCONNECT FABRIC MODULE," filed August 21, 2001; U.S. Provisional Application No. 60/314,287 entitled "INTEGRATED ANALYSIS OF INCOMING DATA TRANSMISSIONS," filed August 22, 2001; U.S. Provisional Application No. 60/314,158 entitled "USING VIRTUAL IDENTIFIERS TO ROUTE TRANSMITTED DATA THROUGH A NETWORK," filed August 21, 2001, and is related to U.S. Patent Application No. _____ entitled "METHOD AND SYSTEM FOR VIRTUAL ADDRESSING IN A COMMUNICATIONS NETWORK," (Attorney Docket No. 030048019US1); U.S Patent Application No. _____ entitled "METHOD AND SYSTEM FOR LABEL TABLE CACHING IN A ROUTING DEVICE," (Attorney Docket No. 030048024US); U.S. Patent Application No. _____ entitled "METHOD

030048039US); U.S. Patent Application No. _____ entitled "USING VIRTUAL IDENTIFIERS TO ROUTE TRANSMITTED DATA THROUGH A NETWORK," (Attorney Docket No. 030048040US); U.S. Patent Application No. _____ entitled "USING VIRTUAL IDENTIFIERS TO PROCESS RECEIVED DATA ROUTED THROUGH A NETWORK," (Attorney Docket No. 030048041US); U.S. Patent Application No. _____ entitled "METHOD AND SYSTEM FOR PERFORMING SECURITY VIA VIRTUAL ADDRESSING IN A COMMUNICATIONS NETWORK," (Attorney Docket No. 030048042US); and U.S. Patent Application No. _____ entitled "METHOD AND SYSTEM FOR PERFORMING SECURITY VIA DE-REGISTRATION IN A COMMUNICATIONS NETWORK" (Attorney Docket No. 030048043US), which are all hereby incorporated by reference in their entirety.

TECHNICAL FIELD

[0002] The described technology relates to a network manager for routing devices of an interconnect fabric.

BACKGROUND

[0003] The Internet has emerged as a critical commerce and communications platform for businesses and consumers worldwide. The dramatic growth in the number of Internet users, coupled with the increased availability of powerful new tools and equipment that enable the development, processing, and distribution of data across the Internet have led to a proliferation of Internet-based applications. These applications include e-commerce, e-mail, electronic file transfers, and online interactive applications. As the number of users of, and uses for, the Internet increases so does the complexity and volume of Internet traffic. According to UUNet, Internet traffic doubles every 100 days. Because of this traffic and its business potential, a growing number of companies are building businesses around the Internet and developing mission-critical business applications to be provided by the Internet.

BRIEF DESCRIPTION OF THE DRAWINGS

[0007] Figure 1 is a network diagram illustrating various nodes of an example Fibre Channel fabric-based interconnect network that are inter-communicating using virtual identifiers.

[0008] Figure 2 is a flow diagram illustrating the discovery processing of a component of the interconnect fabric module in one embodiment.

[0009] Figure 3 is a flow diagram illustrating the discovery processing of the network manager in one embodiment.

[0010] Figure 4 is a flow diagram illustrating the process of establishing a path by the network manager in one embodiment.

[0011] Figure 5 is a flow diagram illustrating the processing of an identify virtual address component of the network manager in one embodiment.

[0012] Figure 6 is a flow diagram illustrating the processing of an initialize label table component of the network manager in one embodiment.

[0013] Figure 7 is a block diagram illustrating a distributed network manager in one embodiment.

[0014] Figure 8 is a flow diagram illustrating the processing of a component of an interconnect fabric module that processes reserved addresses in one embodiment.

DETAILED DESCRIPTION

[0015] A method and system for managing an interconnect fabric that connects nodes is provided. In one embodiment, a network manager manages an interconnect fabric or network of routing devices (e.g., interconnect fabric modules, switches, or routers) to allow source nodes to transmit data to destination nodes. The network manager receives registration requests from source nodes to send data to destination nodes, configures the routing devices of the network to establish a path from each source node to its destination node, and provides a virtual address to each source node. The virtual address identifies a

network manager then sends a query message through the ports of each responding routing device. Alternatively, the network manager can send one query message to the routing device to which it is directly connected and that routing device can forward the query message via each of its ports to the routing device to which it is directly connected. Each port upon receiving the query message may send a message to the network manager with its identification along with the identification of the port to which it is directly connected.

[0020] In one embodiment, each routing device may dynamically discover which of its ports are connected to other devices (e.g., nodes or other routing devices) at initialization. Each port of a routing device may sense a characteristic of its communications link (e.g., voltage on a receive link) or may transmit a request and receive (or not receive) a response via its communications link to identify whether a device is connected. The network manager may poll each routing device for an indication of which ports of the routing device are connected to other devices. The network manager can then send a query message to each connected-to port to identify the port to which it is connected.

[0021] In one embodiment, the network manager establishes paths through the network of routing devices by configuring the ports of the routing devices along the path. The network manager may identify a path from a source node to a destination node using conventional path identification techniques. For example, the network manager may use a shortest path algorithm to identify the path with the smallest number of communications links or may use a congestion-based algorithm that factors in actual or anticipated network traffic to identify the path. The network manager then identifies a virtual address (i.e., a destination virtual address) for the identified path. The virtual address is sent by the source node along with the data to be transmitted to the destination node. The data and virtual address may be stored in a frame (e.g., Fibre Channel or InfiniBand) that has a header and a payload. The header may contain the virtual address and the payload may contain the data. The network manager then configures each source-side port of each routing device along the path to forward frames sent to

the identified virtual address to the destination-side port of the routing device that is connected to the next communications link in the path. The configuration information may be stored in a label table (described below) for the port that maps virtual addresses to destination-side ports. When a source-side port receives a frame with the identified virtual address, it then forwards the frame through the destination-side port in accordance with the configuration information.

[0022] In one embodiment, the network manager identifies a virtual address that is not currently in use by any source-side port along the path. Thus, when a source-side port receives a frame addressed with the identified virtual address, there is no ambiguity as which port of the routing device is the destination-side port. It is possible, however, that paths from two different source nodes to the same destination node may have a common sub-path. For example, the path from one source node may be through communications links A, X, Y, and Z, and the path from the other source node may be through communications links B, X, Y, and Z. In such a case, the network manager may use the same virtual address for both paths and share the terminal portion of the already-configured paths.

[0023] In one embodiment, the network manager may also establish a path between the destination node and the source node. The network manager may identify a new path or may use the same path that was identified between the source node and the destination node (but in the opposite direction). The network manager then identifies a virtual address (i.e., source virtual address) and configures the ports along the path in a manner that is analogous to the configuration of the path from the source node to the destination node. Whenever a source node sends a frame, it may include the source virtual address in the frame. When the destination node receives the frame, it can respond to the source node by sending a frame addressed to the source virtual address.

[0024] In one embodiment, the network manager may need to identify and configure a new path between a source node and a destination node. For example, the network manager may determine that, because of congestion, the required quality of service cannot be provided along the existing path or may

are forwarded to certain manager devices that provide certain functions or services of the network manager.

[0026] In one embodiment, a routing device is an interconnect fabric module ("IFM") with high-speed switching capabilities. An interconnect fabric module can be dynamically configured to interconnect its communications ports so that data can be transmitted through the interconnected ports. Multiple interconnect fabric modules can be connected to form an interconnect fabric through which nodes (e.g., computer systems) can be interconnected. In one embodiment, data is transmitted through the interconnect fabric as frames such as those defined by the Fibre Channel standard. Fibre Channel is defined in ANSI T11 FC-PH, FC-PH-2, FC-PH-3, FC-PI, and FC-FS industry standard documents which are hereby incorporated by reference. One skilled in the art will appreciate, however, that the described techniques can be used with communications standards other than Fibre Channel. In particular, the described techniques can be used with the InfiniBand standard, which is described in the InfiniBand Architecture Specification, Vols. 1-2, Release 1.0, October 24, 2000, which is hereby incorporated by reference. The interconnect fabric module may allow the creation of an interconnect fabric that is especially well suited for interconnecting devices utilizing multiple information types such as might be required by the devices of an enterprise data network ("EDN").

[0027] In one embodiment, a virtual address may be part of a "virtual identifier" (e.g., source or destination identifier) that includes a domain address. A destination identifier of a frame may be set to a virtual identifier. The destination identifiers of the frames received by the interconnect fabric modules are used to forward the frame. Each interconnect fabric module is assigned a domain address. The interconnect fabric modules that are assigned the same domain address are in the same domain. The interconnect fabric modules use of the domain addresses to forward frames between domains. The network manager may configure the interconnect fabric modules with inter-domain paths. When an interconnect fabric module receives a frame with a destination domain address

more resources associated with the destination node, such as an executing application program, a file on storage, or a device that is part of the node. For example, if a source application on a source node initiates a bi-directional communication, a VI NIC for the source node may associate the response virtual identifier with that source application so that received responses can be forwarded to that source application (which then becomes the destination application for those received communications).

[0032] For illustrative purposes, some embodiments are described below in which the VI NIC is used as part of a Fibre Channel or InfiniBand network and/or as part of an EDN architecture. However, those skilled in the art will appreciate that the techniques of the invention can be used in a wide variety of other situations and with other types of networks, and that the invention is not limited to use in Fibre Channel or InfiniBand networks or with EDN architectures.

[0033] Figure 1 is a network diagram illustrating various nodes of an example Fibre Channel fabric-based interconnect network that are inter-communicating using virtual identifiers. In this example embodiment, multiple interconnect fabric modules ("IFMs") 110 with high-speed switching capabilities are used as intermediate routing devices to form an interconnect fabric, and multiple nodes 105, a network manager 115 and a Multi-Protocol Edge Switch ("MPEX") 120 are connected to the fabric. Each of the nodes has at least one VI NIC that uses virtual identifiers when communicating and receiving data. The MPEX is used to connect the Fibre Channel or InfiniBand network to an external network, such as an Ethernet-based network, and similarly includes at least one VI NIC. Data is transmitted through the interconnect fabric using frames such as those defined by the Fibre Channel or InfiniBand standards.

Topology Discovery

[0034] As described above, the network manager may dynamically discover the topology of the network using various different techniques. In the embodiment described below, each interconnect fabric module identifies which of its ports are

interconnect fabric modules have already been selected, then the network manager continues at block 304, else the network manager continues at block 303. In block 303, the network manager retrieves an indication of which ports of the selected interconnect fabric module are connected to other ports. The network manager may send the message using either in-band our out-of-band communications. The network manager then loops to block 301 to select the next interconnect fabric module. In blocks 304-310, the network manager determines the identity of each of the connected-to ports. In block 304, the network manager selects the next interconnect fabric module. In decision block 304, if all the interconnect fabric modules have already been selected, then the network manager completes its discovery process, else the network manager continues at block 306. In blocks 306-310, the network manager loops sending a query message through each port of the selected interconnect fabric module that is connected to another port. In block 306, the network manager selects the next port of the selected interconnect fabric module that is connected to another port. In decision block 307, if all such ports are already selected, then the network manager loops to block 304 to select the next interconnect fabric module, else the network manager continues at block 308. In block 308, the network manager sends a query message through the selected port of the selected interconnect fabric module. In block 309, the network manager receives the identification of the connected-to port of the selected port of the selected interconnect fabric module. The identification may include an indication of the interconnect fabric module and the port number of the connected-to port. In block 310, the network manager stores a mapping between the selected port of the selected interconnect fabric module and the connected-to port of the connected-to interconnect fabric module. These mappings define the topology of the network. The network manager then loops to block 306 to select the next port of the selected interconnect fabric module that is connected to another device.

[0036] The processing of the discovery of the network manager as described above assumes that the network manager initially is aware of all interconnect

fabric modules of the interconnect fabric. One skilled in the art will appreciate that the network manager may become of aware additional interconnect fabric modules during the discovery process. For example, if the network manager is centralized, then it may initially send a query message through its port that is connected to the interconnect fabric. The receiving port responds with the identity and interconnect fabric module and its port number. The network manager can then requested that identified the interconnect fabric module to provide a indication of which of its ports are connected to other ports. The network manager can then send a query message through each of the indicated ports to the connected-to ports. The connected-to ports then respond with the identification of the connected-to interconnect fabric module and connected-to port. This process can be repeated transitively by the network manager to identify all interconnect fabric modules that comprise the interconnect fabric.

Establishing a Path

[0037] Figure 4 is a flow diagram illustrating the process of establishing a path by the network manager in one embodiment. A path is typically established when a node registers with the network manager. An establish path component of the network manager may receive an indication of a source node and a destination and then identify paths of ports of interconnect fabric modules from the source node to the destination node and from the destination node to the source node. The component then identifies virtual addresses for the paths and initializes the label tables of the ports of the interconnect fabric modules along the identified paths. A label table of a port contains mappings from virtual addresses to destination-side ports through which a frame sent to that virtual address is to be forwarded. In block 401, the component identifies the paths. In one embodiment, the path from the source node to the destination node and the path from the destination node to the source node use the same ports of the same interconnect fabric modules. That is, the paths use the same communications links. Alternatively, the path in one direction may be different from the path in the other

is available. The virtual address may not be available to a port along the path when that port already uses that virtual address. In blocks 501, the component selects to the next virtual address. In decision block 502, if all the virtual addresses have already been selected, then the component indicates that a virtual address could not be identified, else the component continues at block 503. In blocks 503-505, the component loops selecting each port along the path and determining whether that port already uses the selected virtual address. In block 503, the component selects the next interconnect fabric module and port of the path. In decision block 504, if all the interconnect fabric modules and ports of the path have already been selected, then the component uses the selected virtual address as the identified virtual address and then completes, else the component continues at block 505. In decision block 505, if the selected virtual address is available at the selected interconnect fabric module and selected port, then the component loops to block 503 to select the next port along the path, else the component loops to block 501 to select the next virtual address.

[0039] Figure 6 is a flow diagram illustrating the processing of an initialize label table component of the network manager in one embodiment. The initialize label table component sends a command to each port along the path indicating to add a mapping from the identified virtual address to the other port of that interconnect fabric module along the path. The component is passed in indication of the path, the virtual address, and an indication of whether the virtual address is a source virtual address or a destination virtual address. In block 601, the component selects the next interconnect fabric module and port in the path based on whether the source or destination virtual address has been passed. In decision block 602, if all the interconnect fabric modules along the path have already been selected, then the component completes, else the component continues at block 603. In block 603, the component sends a message to be selected port of the interconnect fabric module indicating to add to its label table a mapping from the virtual address to the other port of the path. The component then loops to block 601 to select the next interconnect fabric module and port in the path.

embodiment. This component forwards the frame to the network manager via either in-band or out-of-band communications. With the use of in-band communications the frame can be routed to the appropriate interconnect fabric module, which can then send the frame to the network manager using the out-of-band communications. In block 801, if the virtual address of the received frame is a reserved address, then the component continues at block 802, else the component completes. In decision block 802, if the virtual address parameter within the frame is in the label table, then the frame is to be forwarded using in-band communications and the component continues at block 804, else the frame is to be forwarded directly to the network manager at the IFM's manager device using out-of-band communications and the component continues at block 803. In block 803, the component forwards frame to the administrative port and then completes. In block 804, the component forwards the frame based on the port map of the label table and then completes.

[0043] One skilled in the art will appreciate that, although various embodiments of the technology have been described, various modifications may be made without deviating from the spirit and scope of the invention. Accordingly, the invention is not limited except as by the appended claims.